2. Quais são as etapas mais comuns de um pipeline de bioinformática de NGS para DNAseq

e quais ferramentas podem ser usadas em cada etapa?

Podemos dividir uma pipeline de bioinformática de NGS para DNAseq em quatro etapas:

1. Análise de qualidade do sequenciamento;
2. Alinhamento e análise de qualidade do Alinhamento;
3. Chamada de variantes;
4. Anotação de variantes;
5. Análise de qualidade do sequenciamento: Nessa etapa avaliamos os arquivos FASTQ que recebemos do sequenciamento e também realizamos o processo de trimagem para remover a região de baixa qualidade com alto risco de ser erro de sequenciamento e de primers e adaptadores.

Algumas ferramentas que podem ser usadas são:

* fastqc: Para analisar e visualizar os arquivos fastq
* cutadapt e trimmomatic : Processo de trimagem

1. Alinhamento e análise de qualidade do Alinhamento: O alinhamento (ou mapeamento) de leituras em um genoma de referência é uma etapa vital para analisarmos possíveis variantes, sejam elas variantes de um único nucleotídeo (SNV) ou outros tipos de variantes. Leituras podem ser mapeadas (ou alinhadas) em relação a uma sequência conhecida e de qualidade - o genoma de referência - para que possamos comparar nossa amostra e encontrar diferenças - ou variantes.

Algumas ferramentas que podem ser usadas são:

* bwa: Alinhamento contra um genoma de referência (output arquivo sam)
* Picard: Avaliar o alinhamento e melhorar sua qualidade (remover as duplicações do alinhamento)
* samtools: visualização do alinhamento e conversão do arquivo sam para bam.

1. Chamada de variantes: Após a preparação dos arquivos SAM e BAM, podemos executar a chamada de variantes. Chamada de variante é o processo pelo qual identificamos variantes de dados de sequência. Nessa etapa identificamos onde as leituras alinhadas diferem do genoma de referência e gravamos em um arquivo VCF.

Algumas ferramentas que podem ser usadas são:

* Bcftools, GATK: Chamada de variantes

1. Anotação de variantes: A anotação de variantes é o processo de atribuir informações funcionais a variantes de DNA. Existem muitos tipos diferentes de informações que podem ser associadas a variantes, desde medidas de conservação de sequência até previsões sobre o efeito de uma variante na estrutura e função da proteína.

Algumas ferramentas que podem ser usadas são:

* ANNOVAR: Anotação de variantes através de arquivos VCF e GFF.